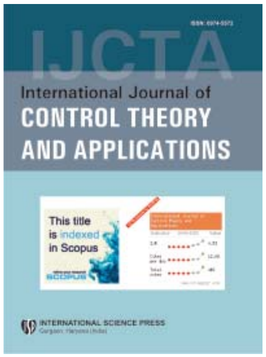

International Journal of Control Theory and Applications

ISSN : 0974–5572

© International Science Press

Volume 10 • Number 23 • 2017

Analyzing Customer Behavior Through Segmentation using Data Mining Techniques

S. Kavitha[a] and S. Manikandan[b]

[a]Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore, TamilNadu, India.

E-mail: kavithapandian.s@gmail.com

[b]Prof and Head, Department of CSE, Sriram Engineering College, Chennai, TamilNadu, India.

E-mail: manidindigul@rediffmail.com

Abstract: E-Commerce is a Killer-domain for data mining. Data mining is the automate detection of relevant pattern from the database. E-Commerce is a very famous as well as frequently used new technique in the real world applications. Data mining (DM) is a collection of exploration techniques. Which is based on advanced analytical methods and tools. DM also used for handling large volume of information. Using DM tools we can predict the behaviors and trends of future. It allowing businesses to make upbeat way for the customer. In this research work, it is taken online shoppers purchasing vehicle data set and find accuracy in terms of its purchasing behavior using some of the classification algorithms. The classification algorithms namely Bayesian belief network and Navie Bayes are utilized for the analysis and a comparative study of both the algorithms are carried out. Finally, the performance of the chosen algorithm is suggested for analyzing the vehicle data set based on the purchasing behavior of the customer and predicts some accuracy.

Keywords: *Classification Algorithms, Bayes Net, Naïve Bayes Algorithm.*

1. INTRODUCTION

The major challenging task of businesses is to understand the needs of customer, we are not able to conclude or predict their buying behavior. One of the way is we can predict from their past purchases otherwise no way to predict them, this kind of prediction helps the businesses to attain and retain the customers. Try to involve the customer for the checking the quality, fixing price, because they are the potential customers, so we must retain the customers, so keep in touch with always, sometime announcing incentives, gifts gift voucher for the occasion periods. Invite the selective customers into plant for visiting the product manufacturing process which is one of the way of advertising products. The first purchase is not the last one, so bringing them into long term relationship.

Jayendra Sinha and Jiyeon Kim are studied about the insights of online retailing in India – specifically factors affecting Indian consumers' online buying behavior[1]. Although the convenience risk seemed to be the only factor significantly affecting Indian consumers' online purchases, when looking at male and female perceptions, there were different factors affecting male/female consumer's behaviors. Perceived risk is significant for male but not for female. Dr. Sankar Rajagopal[2] identification of high-profit, high-value and low-risk customers via the data mining technique customer clustering has been studied using IBM Intelligent Miner. [3]Sujatha Joshi, Abhijit Chi Putkar and Yatin Jog they have been analyzed branding factors play an important role in influencing customer to retain loyal into their brand. From the analysis married couples visit the store more often than unmarried persons, customers are visiting the stores based on word of mouth marketing rather than other means of marketing[4].

Aditya Kumar Gupta & Chakit Gupta[5] the main contribution of his paper lies in the focusing important issues to improve decision making to optimize your relationships with Customer in highly Customer based business. The study identifies major problems of Customer behavior and that has direct impact on the sale and production. According to the author the changes in the technology has changed the way people form an attitude and purchase intention rising trend in wide range of users[6]. The remaining papers is structured as follows. Section II discusses about the materials and methods used for this research work. The experimental results are explained in section III. Finally, section IV concludes the research work.

2. MATERIALS AND METHODS

Classification which means that the estimation of association or we called as a separation or ordering of objects into classes. Let we see the brief introduction about various algorithm for using classifications.

2.1. Naive Bayesian Classifier

Figure.1 shows the structure of a Naive Bayes classifier, which is the simplest form of useful Bayesian network classifier. (We use the term useful, as actually this is not the simplest possible classifier. We could perform classification using a model with no links, however a model of this type would have limited use). The links in a Naive Bayes model are directed from output to input, which gives the model its simplicity, as there are no interactions between the inputs, except indirectly via the output. Note however that directing links from output to input, is not a requirement for all Bayesian network classifiers.

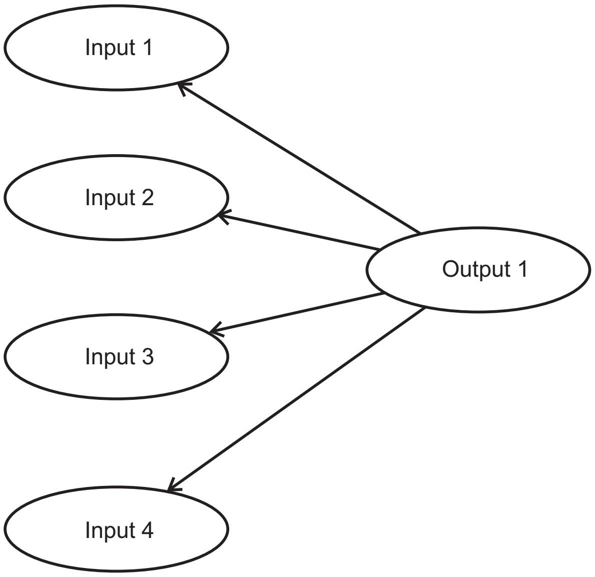

Figure 1: Naïve Bayesian Classifier

NAIVE based method is based on the work of Thomas Bayes (1702-1761).Bayes was a British Minister and his theory was published only after his death. This theorem has some notations which is defined as

P(A) – Refers to the probability that event A will occur.

P(A/B) – Refers for the probability that event A will happen, given that event B already happened.

The Bayes' theorem:

$$P(A / B) \;=\; P(B / A)\, P(A)\, P(B)$$

The probability belonging to one of the Classes C1, C21, C3 etc by calculating $P(C_i / X)$. Once these probabilities have been computed for all these classes, we simply assign X to the class that has higher conditional probability. Now let us see how $P(C_i / X)$ is calculated.

We have $P(C_i / X) \;=\; [P\,(X / C_i) / P(C_i)] / P(X)$

$P(C_i / X)$ – Probability of the object X belonging to class C_i.

$P(X/C_i)$ – Probability of obtaining attribute values X, if we know that it belongs to C_i.

$P(C_i)$ – Probability of any objects belonging to class C_i. Without any other information.

2.2. Bayesian Belief Network

Bayesian networks are widely used to perform classification tasks, with the following advantages. It is based on the probability theory, and it is not a black box approach which allow rich in structure. It can mix expert opinion and data to build models. Additionally, which is used to predicting outputs from the given inputs. Support for missing data during learning and classification.

A belief networks is defined by two components such as,

1. A directed acyclic group
2. And a set of conditional probability tables.

It simplifies the computation, and it is most accurate when compare with other classifiers.

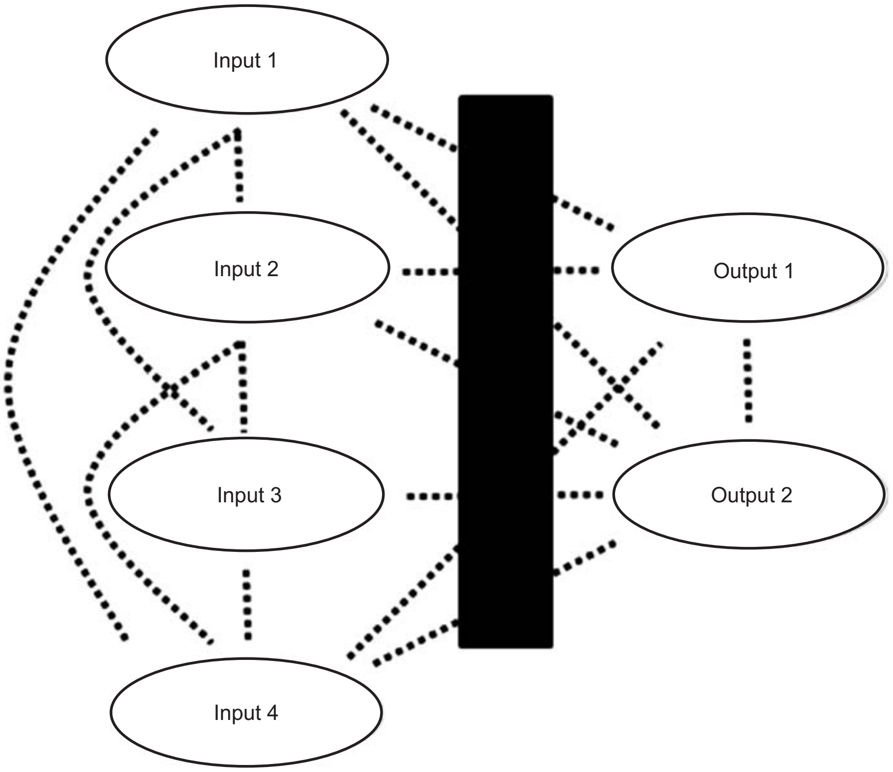

Figure 2: Bayesian Belief Network classifier

Figure 2 depicts the possible structure of a Bayesian network used for classification. The dotted lines denote potential links, and the blue box is used to indicate that additional nodes and links can be added to the model, usually between the input and output nodes. In the learning and training of a belief network, a number of scenarios is possible. The network topology may be given in advance or inferred from the data. The network variables may be observable or hidden in all or some of the training tuples. The case hidden data is also referred to as *missing values* or *incomplete* data.

2.3. Statistical Measures

The various formulas used for the calculation of different measures are as follows. The following formula is used to calculate the proportion of the predicted positive cases, Precision P using TP is True Positive Rate and FP is False Positive Rate and they defined as,

$$\text{Precision P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

It has been defined that Recall or Sensitivity or True Positive Rate (TPR) means the proportion of positive cases that were correctly identified. It will be computed as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Where FN = False Negative Rate

In this research work, it is carried out the following three measures namely correctly classified instances, incorrectly classified instances, and accuracy. Correctly classified instances are the instances which are correctly classified by any classification algorithm. Percentage of correctly classified instances is called as accuracy. Incorrectly classified instances are the instances that they are not correctly classified by the algorithm. Sometimes it is observed that the data which is incorrectly classified may contain inconsistent data, noisy or data out of scope.

Accuracy is calculated by a measured value which is closed to the true value.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

The above formula will calculate the accuracy (the proportion of the total number of predictions that were correct) with TN = True Negative. The sensitivity is the percentage of positive records classified correctly out of all positive records.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{4}$$

The specificity is the percentage of positive records classified correctly out of all positive records.

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{5}$$

The F-Measure can be computed as some average of the information retrieval precision and recall metrics.

$$\text{F} = \frac{2^{*}\ \text{Recall} * \text{Precision}}{\text{Precision} + \text{Recall}} \tag{6}$$

Kappa Statistics measure degree of agreement between two sets of categorized data. Kappa result varies between 0 to 1 intervals. Higher the value of Kappa means stronger the agreement. Kappa is a normalized value of agreement for chance of agreement.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \qquad (7)$$

Where P(A) = percentage of agreement, P(E) = chance of agreement. If K = 1 agreement is perfect between the classifier and ground truth. If K = 0 indicates there is a chance of agreement. The mean absolute error (MAE) is a quantity used to measure predictions of the eventual outcomes. The mean absolute error is given by

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n} |e_i| \qquad (8)$$

The mean absolute error is an average of the absolute errors $e_i = |f_i - y_i|$, Where f_i = prediction, y_i = true value.

Root mean squared error is the square root of the mean of the squares of the values. It squares the errors before they are averaged and RMSE gives a relatively high weight to large errors. The RMSE Ei of an individual program *i* is evaluated by the equation:

$$Ei = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \left(P(i,j) - Tj\right)^2} \qquad (9)$$

Where, P(*i*, *j*) = the value predicted by the individual program, *i* = fitness case, T*j* = the target value for fitness case *j*. ROC Area is defined as area under the ROC curve which is the probability of randomly chosen positive instance that is ranked above randomly chosen negative one. Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values. Mathematically, the relative absolute error E_i of an individual program is evaluated by the Equation:

$$E = \sum_{j=1}^{n} |Pij - Tj| \Big/ \sum_{j=1}^{n} |Tj - T''j| \qquad (10)$$

Where, P_{ij} *s* the value predicted by the individual pro- gram *i* for sample case *j* (out of n sample cases); T_{ji} *s* the target value for sample case *j* ; and T_i *s* given by the formula:

$$T'' = \frac{1}{n}\sum_{j=1}^{n} Tj \qquad (11)$$

Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted values.

3. EXPERIMENTAL RESULTS

Table 1
Description of the Data Set

S.No.	*Description*	*Value*
1.	Gender	Female/Male
2.	Location	Rural/Suburban/Urban
3.	Employment Status	Employed/Unemployed
4.	Marital Status	Single/Married/Divorce
5.	Vehicle Class	Two door/four door/SUV/Luxury SUV/Sports Car/Luxury Car
6.	Vehicle Size	Small/Med size/Large

Online shoppers purchasing vehicle data set having 20 attributes and 9134 instances. This research work mainly discusses about the accuracy of classification algorithms compared with the execution time and error rate using WEKA software. From the results of classification Naive Bayes was classified better than the Byes Belief Network. Bayes' theorem assumes that all attributes are independent and that the training sample is good sample to estimate probabilities. These assumptions are not always true in practice, as attributes are often correlated but in spite of this the Naive Bayes method performs reasonably well. Other techniques have been designed to overcome this limitation. One approach is to use Bayesian networks that combine Bayesian reasoning with casual relationships with attributes.

In this data set I was taken 6 attributes from the vehicle purchase database, these are the specific attributes purchaser those who are from rural, suburban, urban, and the employment status namely employed and unemployed, the major three classes of the data set is small, medium, large vehicle size.

Table 2
Results of two measures

Algorithms	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*ROC Area*	*Class*
Bayes Net	0.848	0.777	0.721	0.848	0.78	0.604	Med size
	0.236	0.154	0.268	0.236	0.251	0.632	Small
	0.011	0.001	0.455	0.011	0.021	0.54	Large
	0.643	0.577	0.606	0.643	0.599	0.603	Weighted Average
Naves Bayes	0.956	0.951	0.704	0.956	0.811	0.593	Med size
	0.046	0.041	0.211	0.046	0.075	0.626	Small
	0.003	0.004	0.086	0.003	0.006	0.547	Large
	0.681	0.677	0.545	0.681	0.586	0.595	Weighted Average

Total Number of Instances : 9134

Table 3
Error Reports

Statics	*Bayes Net*	*Navie Bayesd*
Kappa statistic	0.0738	0.0054
Mean absolute error	0.337	0.3227
Root mean squared error	0.4022	0.3948
Root mean squared error	0.4022	0.3948
Relative absolute error	110.5228 %	105.8252 %
Root relative squared error	103.0032 %	101.1048 %

The accuracy is calculated based on addition of true positive and true negative followed by the division of all possibilities. Accuracy is measured and created using 10 fold cross validation method. Tenfold cross-validation is the standard way of measuring the error rate of a learning scheme on a particular dataset for reliable results, 10 times the data set is executed by 10-fold cross-validation. In 10-fold cross validation method, the data set is randomly sub divided into ten equal sized partitions.

Among the partitions nine of them are used as training set and the remaining one is used as a test set. Evaluation of performance is compared using Mean absolute error, Root mean squared Error, Receiver Operating Characteristic (ROC) Area and Kappa statistics. Large test sets gives a good assessment of the classifier's performance and small training sets which result in a poor classifier by using this method. The Table 4 gives the error values of the taken two classification algorithms.

Table 4
Performance Accuracy

Algorithms	*Correctly Classified Instances*	*Incorrectly Classified Instances*
Bayes Net	64.342 %	35.658 %
Naive Bayes	68.1301 %	31.8699 %

Classify the shoppers data correctly from the training data set, the error rates and accuracy are calculated using classifiers. The accuracy of Bayes Net algorithm found to be 64.34%, Naive Bayes is 68.13%,.The results of various measures are given table 2. From those classification Med size car preferred 70.33% of people, small size car 18.98% of them preferred, the rest of large size car preferred 10.35% of people.

4. CONCLUSIONS

In this research work, it is discussed the classification techniques such as Naïve Bayes classifier algorithm, Bayesian belief networks algorithm are separating the objects into classes. It was analyzed by conducting the experiments using the marketing analysis data set. The performance is known by the indicators such as accuracy, specificity, precision, and error rate. The data preprocessing method filter the missing data and noise data. For processing I used Naïve Bayes method. The data preprocessing method can improve the accuracy of the classifier because it removes the noise data or the missing values. To improve the overall accuracy use large data set and increase the number of cross fold validation.

5. REFERENCES

[1] Jayendra Sinha (USA), Jiyeon Kim (USA) "Factors affecting Indian consumers' online buying behavior", Innovative Marketing, 8(2), 2012, p.

[2] Dr. Sankar Rajagopal, Enterprise DW/BI Consultant ,Tata Consultancy Services, Newark, DE, USA, "Customer Data Clustering Using Data Mining Technique", International Journal Of Database Management Systems (Ijdms) 3(4) 2011, pp

[3] Sujatha Joshi, Abhijit Chi Putkar and Yatin Jog, "Influence of Brand Oriented Factors on Customer Loyalty of Prepaid Model Services", Indian Journal of Science and Technology,8(S6), 2015,43-49.

[4] K.Kannan and K.Raja, "Decision Making Process for B2C Model Using Behavior Analysis with Big Data Technologies". Indian Journal of Science and Technology, 9(24), 2016.

[5] R.Deiva Veeralakshmi, "A study on online shopping behavior of customers", International journal of scientific research and management (ijsrm) ISSN (e): 2321-3418.

[6] E.W.T. Ngai ,*, Li Xiu , D.C.K. Chau, "Application of data mining techniques in customer relationship management:" journal homepage: www.elsevier.com/locate/eswa.

[7] Aditya Kumar Gupta & Chakit Gupta, "Analyzing customer behavior using data mining Techniques: optimizing relationships with customer" International Journal Of Management Insight, 6(1),2010 pp.

[8] Prabha Kiran and S.Vasantha, "Transformation of Consumer Attitude through Social Media towards Purchase Intention of Cars", 9(21), 2016. P7.

[9] Krishna R.Kashwan, Member of IACSIT, and C.M.Velu, "Customer Segmentation using Data mining Techniques" 5(6) 2013, pp.

[10] Dattatray V.Bhate, M.Yaseen Pasha, “Analyzing target customer behavior using datamining techniques for e-com”.

[11] N.R.Srinivasa Raghavan, “Data mining in e-commerce”, Sadhana, 3(2), 2005.pp.

[12] Mohammad Ali Farajian, Shahriar Mohammadi, “Mining the Banking Customer Behavior Using Clustering and Association Rules Methods”, International Journal of Industrial Engineering & Production Research, 21(4), 2010 pp.. 239-245.

[13] Belsare Satish and Patil Sunil, “Study and Evaluation of user’s behavior in e-commerce Using Data Mining”, www.isca.in.

[14] Jayendra Sinha (USA), Jiyeon Kim (USA), “Factors affecting Indian consumers’ online buying behavior”, Innovative Marketing, 8(2), 2012, pp.

[15] Abdullah N, Xu Y, Geva S, Chen J., “Infrequent purchased product recommendation making based on user behavior and opinions in e-commerce sites”. 2010 IEEE International conference on Data Mining Workshops, ICDMW, Sydney, NSW,2010, p 1084-91.

[16] Kaferer JN, “The New Strategic Brand Management London:” Kogan Page:2008.

[17] Doolin B, Dillon.S, Thomson F, Corner H, Perceived Risk, “The Internet Shopping Experience and online Purchasing Behavior:” A New Zealand perspective. Journal of Interactive Advertising.2009:10(1):77-93.

[18] Schau HJ, Munish AM, Arnould EJ. “How brand community practices create value”, Journal of Marketing Science.2009:34(2):115-27.